**Taaltoetsen ontwikkelen: tips en valkuilen**
Bart Deygers

## Inleiding: taaltoetsen in het onderwijs

In een recent artikel over toetsgeletterdheid bij taalleerkrachten stelt Fulcher dat "the first decade of the 21st century has seen a phenomenal increase in the testing and assessment responsibilities placed upon language teachers"(Fulcher 2012: 113). Zo heeft de komst van het ERK (Common European Framework of Reference for Languages) in 2001 een onweerlegbare invloed gehad op de praktijk van taaltoetsen (McNamara & Roever 2006). Onderwijsoverheden zagen en zien in het ERK een handig instrument om uitstroomniveaus voor taalvakken te bepalen, hoewel het een uitgesproken descriptieve, geen normatieve, inslag heeft (Jones 2011). In Nederland heeft het "ERK Masterplan"de invloed van het ERK op het taalonderwijs en taaltoetsen bevestigd en ook in Vlaanderen zijn de eindtermen voor vreemde talen aan het ERK gekoppeld. Een en ander heeft verregaande gevolgen voor de leerkracht. Toetsen moeten niet meer louter aansluiten bij het geleerde, maar moeten ook een adequate maat zijn voor het beoogde ERK-niveau. Zeker in een Vlaamse context, waar er geen traditie van gecentraliseerd toetsen heerst, maakt dat van de leerkracht een belangrijke toetsontwikkelaar.

De taalleerkracht anno 2012 wordt niet enkel verondersteld om niveautoetsen te kunnen opstellen; er wordt ook van hem of haar verwacht om *veel* toetsen op te stellen. Van de zes jaren die een Vlaamse ASO-leerling op school doorbrengt, gaat er in totaal één volledig jaar op aan toetsen (Struyf 2004). Dat hoeft niet noodzakelijk een slechte zaak te zijn, zolang er op een goede manier getoetst wordt. Zo stelt de dynamic assessment beweging dat frequent toetsen de motivatie aanzwengelt en een cultuur van succes promoot (Lantolf, 2009). Ook binnen de logica van breed evalueren past het idee van veelvuldig evalueren, zolang het gebeurt op diverse manieren en in diverse contexten.

Evalueren waar je staat ten opzichte van waar je komt en waar je heen wil, is een essentieel deel van het leerproces (Shepard 2000). Leerlingen vaak toetsen is niet noodzakelijk slecht, maar leerlingen vaak *slecht* toetsen is onmiskenbaar nefast. Toch krijgen taalleerkrachten onvoldoende opleiding en ondersteuning bij het opstellen van toetsen en is niet elke taalleerkracht even toetsgeletterd (Fulcher 2012). Deze bijdrage wil dan ook een praktische en concrete aanzet tot taaltoetsen zijn; een overzicht van essentiële toetsprincipes, van gevaarlijke valkuilen en van tips om die valkuilen te vermijden.

## Taaltoetsen 1.1: basisvragen

Alles begint met een construct: "some abstract belief of what language is, what language profiency consists of, what language learning involves and what language learners do with language " (Alderson, Clapham & Wall 1996: 16). Het construct van een toetsontwikkelaar –

zijn of haar visie op taal – heeft een zeer grote invloed op de inhoud en de vorm van een taaltoets. Als de toetsontwikkelaar taal beschouwt als een som van verschillende subvaardigheden en kennisonderdelen, zal de toets er anders uitzien dan wanneer de toets ontwikkeld is door iemand die taal ziet als een geïntegreerd geheel. Voor je een toets gaat ontwikkelen, moet je dus jouw construct expliciteren. Pas daarna kan je de basisvragen van toetsontwikkeling behandelen.

Het is belangrijk om als leerkracht stil te staan bij het *waarom* van toetsen. De verschillende redenen om te toetsen verraden immers ook de impact van een toets. De ene toets is immers belangrijker dan de andere in het leven van een student. Een taaltoets voor migratiedoeleinden kan een fundamentele impact hebben op de levensloop van een kandidaat en is dus relatief gezien belangrijker dan een motivationele toets die talige vooruitgang meet. In de literatuur wordt het onderscheid tussen levensbepalende toetsen en toetsen met een beperkte impact aangeduid met *high stakes* en *low stakes* (Chapelle et al. 2003). Het spreekt vanzelf dat levensbepalende toetsen aan striktere kwaliteitscriteria moeten voldoen dan toetsen met een beperkte impact.

De meest essentiële vraag die je je als leerkracht stelt bij het opstellen van een toets is dus niet “Wat zal ik toetsen?”. Op die vraag kan je pas een antwoord geven wanneer je weet waarom je toetst (Brindley 2001, Inbar-Louise 2008, Bachman 2010). Een plaatsingstoets bijvoorbeeld heeft een heel andere inhoud dan een toets die aan het eind van een curriculum komt omdat de reden om te toetsen verschilt. Terwijl de eerste toets een curriculumonafhankelijke inhoud heeft, kan de tweede best behoorlijk aansluiten bij de behandelde leerstof.

De toetsinhoud vormt het antwoord op de tweede vraag: “Wat?”. Zoals hierboven gesteld, wordt de inhoud van een toets mee bepaald door de reden om te toetsen. Diagnostische taaltoetsen hebben een andere, bredere, inhoud dan toetsen die peilen naar kennis van een grammaticale constructie. Het is uiterst belangrijk om een duidelijk en helder idee te hebben over wat je wil toetsen. Het meest essentiële kwaliteitscriterium van een taaltoets – validiteit – is er immers nauw mee verbonden.

Betrouwbaarheid en haalbaarheid, twee andere kwaliteitscriteria, hangen samen met de derde vraag: hoe toetsen. Er zijn tientallen manieren om een toets af te nemen. Je kunt open of gesloten vragen gebruiken, je kunt direct of indirect toetsen, geïntegreerd of discreet en ga zo maar door. De veelheid van mogelijkheden wordt echter beperkt door het gekozen construct en door de antwoorden op de waarom- en de wat-vragen.

De tabel hieronder geeft een voorbeeld van hoe toetsontwikkelaars met een verschillend construct, maar een gelijklopend doel zouden kunnen omgaan met de basisvragen van

toetsontwikkeling. Het gaat hier natuurlijk om een fictieve voorstelling van twee extremen op een continuüm. Er zijn vele tussenvormen mogelijk.

| | **Toets 1** | **Toets 2** |
|---|---|---|
| **Construct** | Elementgericht.<br>Taal bestaat uit verschillende elementen: de vier vaardigheden, aangevuld met lexicale en grammaticale kennis. | Geïntegreerd.<br>Taal wordt nooit vrij van context of sociaal gebruikt en taal valt niet op te splitsen in gescheiden elementen. |
| **Waarom?** | Niveaubepaling.<br>Ik wil weten hoe goed mijn nieuwe leerlingen zijn en waar hun zwakke punten liggen. Ik wil te toets gebruiken als eerste meting om de vooruitgang van mijn leerlingen te kunnen opvolgen. Het is dus een low stakes toets. | |
| **Wat?** | Grammatica, woordenschat en (een selectie uit) de vier vaardigheden. | Ik toets de vaardigheden zo veel mogelijk in een authentieke context en combineer ze waar mogelijk. |
| **Hoe?** | Ik geef een toets die bestaat uit algemene woordenschat en grammatica die mijn nieuwe leerlingen zouden moeten beheersen. Ik vraag ze ook een tekst te schrijven. Lees- en luistervaardigheid test ik via meerkeuzevragen. | Mijn leerlingen schrijven een tekst op basis van een radioreportage die bij hun leefwereld aansluit interesseert. Hun niveau van grammatica en woordenschat leid ik af uit hun tekst. Ik laat de leerlingen ook mondeling rapporteren over een gelezen tekst. |

Tabel 1

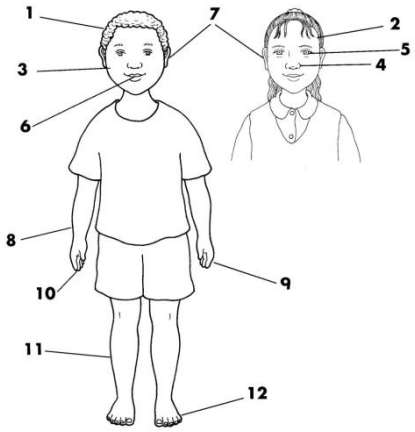
**Taaltoetsen 1.2: Kwaliteitscriteria**

Een toets meet altijd iets, maar “iets” is niet altijd wat je wil meten. In een ideale taaltoets spelen twee variabelen mee: de moeilijkheid van de taken en de vaardigheid van de kandidaat. In de realiteit meet een toets ook heel wat ruis. Zo kan een digitale toets een kandidaat met beperkte computervaardigheden benadelen en iemand met een slecht handschrift bevoordelen. Maar ook schijnbaar modale gebeurtenissen zoals straatlawaai, hoofdpijn of een ochtendhumeur kunnen een score negatief beïnvloeden, ook al hebben ze niets met taalvaardigheid te maken.

Omdat een toets altijd een stukje ruis meet, is het belangrijk om op verschillende manieren te toetsen. Toetsen die afgenomen worden in een strikt formele context kunnen stressgevoelige leerlingen benadelen waardoor ze onderpresteren (Cassady & Johnson 2002). Als je een heel jaar lang formele toetsen zou afnemen, zou je het effect van die stressgevoeligheid disproportioneel versterken. Het is dus essentieel om gevarieerd te

toetsen, maar zelfs wie gevarieerd toetst, moet enkele belangrijke kwaliteitscriteria in acht nemen: validiteit, betrouwbaarheid en haalbaarheid.

Een valide toets meet wat vooropgesteld werd te meten. Dat klinkt vanzelfsprekend, maar dat is het niet. Zo meet de afbeelding links een zeer specifiek lexicaal veld. Het zou verkeerd zijn om op basis van deze taak conclusies te trekken over het bereik van iemands woordenschat. Meten wat je wil meten is de essentie van toetsen (Bachman 2000, Alderson & Banerjee 2002, McNamara 2006), ook al lijkt dit een louter theoretische classificatiekwestie. Stel: de elementgerichte leerkracht uit de tabel hierboven wenst het hele jaar lang de vier vaardigheden apart te toetsen, maar hij of zij toetst elke receptieve vaardigheid telkens op een productieve manier (bv door een schrijfopdracht te koppelen aan een leesopdracht). In dat geval wordt er een jaar lang zeer beperkt – en dus fout – getoetst en weet de leerkracht eigenlijk niets over de luister- en leesvaardigheid van de leerlingen. Geen enkele vaardigheid is immers "zuiver gemeten"; elke meting werd vertroebeld doordat een productieve vaardigheid gebruikt werd als middel om een receptieve vaardigheid te meten. Dit alles betekent niet dat het af te raden zou zijn om receptieve en productieve taken te mengen. Het betekent wel dat je als leerkracht perfect moet beseffen wat je wil meten en dat je kritisch dient stil te staan bij hoe je meet wat je wil meten.




Validiteit is het fundament van een kwaliteitsvolle toets is, maar het betekent niets als die toets onbetrouwbaar is. Betrouwbaarheid heeft betrekking op de consistentie van het meetinstrument. Anders gezegd: stel dat lerares Sofie morgen een toets afneemt van leerling Lisa. Sofie oordeelt dat Lisa geslaagd is met een score van 8 op 10 en stopt de toets een jaar lang weg. Na een jaar bekijkt Sofie de toets opnieuw. Als Lisa nu plots een heel andere score, kunnen we twijfelen aan de betrouwbaarheid van de toets, de beoordelingsprocedure of de beoordelaar.

Het is duidelijk dat de betrouwbaarheidseisen die aan een toets gesteld worden recht evenredig zijn met de *high stakeness* ervan. Hoe meer impact een toets heeft op het leven van een kandidaat, hoe eerlijker en consistenter dat oordeel moet zijn. Grote internationale taaltoetsen houden de variabelen die de betrouwbaarheid van een toets kunnen beïnvloeden grondig in de gaten: onbetrouwbare items worden verwijderd, de mildheid of strengheid van een beoordelaar wordt bijgesteld bij de berekening van de eindscore etc. De betrouwbaarheid van een toets controleren hoeft echter niet steeds gepaard te gaan met grootschalige pretests en statistische formules. Zo helpt het consequent gebruik van een beoordelingsmodel de betrouwbaarheid van de beoordeling van productieve vaardigheden al een groot stuk vooruit (Knoch 2009). Wanneer verschillende beoordelaars eenzelfde

beoordelingsmodel gebruiken om dezelfde toets te beoordelen, valt het ten zeerste aan te raden om die beoordelaars te trainen, zodat ze allemaal dezelfde invulling geven aan de criteria (Weigle 1994, Lumley 2002). Duidelijk opgestelde instructies en criteria, hebben eveneens een groot effect op de betrouwbaarheid van een toets.

Een derde en laatste kwaliteitscriterium dat in deze tekst belicht wordt, is haalbaarheid. Een haalbare toets kan binnen een aanvaardbare tijdspanne afgenomen worden op een manier die geen overdreven technische of praktische belasting vraagt. Rekening houden met haalbaarheid is dus rekening houden met werkbelasting en de technische mogelijkheden van de plaats waar je een toets afneemt. Stel: je wil een digitale schrijftoets afnemen op basis van een online videofragment, maar het netwerk van je school kan die extra belasting niet aan. Je schakelt op het laatste moment dus noodgedwongen over op een papieren toets. Daardoor meet je niet meer wat je wil meten (validiteit) en gaan er misschien andere criteria meespelen dan je oorspronkelijk in gedachten had (betrouwbaarheid).

Haalbaarheid hangt dus samen met validiteit en betrouwbaarheid en validiteit betekent pas iets als de toets ook betrouwbaar is. De drie kwaliteitscriteria zijn dus onlosmakelijk met elkaar en met de drie basisvragen verbonden. Door consequent de juiste vragen te stellen zij de toetsontwikkeling, kan je dus met een minimum aan inspanning kwalitatief hoogstaande toetsen ontwikkelen.

| **Basisvraag** | **Kwaliteitscriterium** |
|---|---|
| *Waarom* meet ik? | Mijn kwaliteitseisen moeten strenger zijn bij een *high stakes* toets. |
| *Wat* meet ik? | Meet ik wat ik wil meten? |
| *Hoe* meet ik? | Meet ik betrouwbaar?<br>Is mijn toets haalbaar? |

Tabel 2

**Slot: taalleerkracht of toetsexpert?**

Het doel van deze bijdrage is niet om elke leerkracht Nederlands om te vormen tot een psychometrisch tovenaar of een toetsexpert. De taak van een leerkracht is in de eerste plaats didactisch – gelukkig maar.
Maar als er dan toch moet getoetst worden, kan dat maar beter goed gebeuren. Leerlingen zowel als leerkrachten hebben immers baat bij een valide en betrouwbare inschatting van taalvaardigheid.

**Bibliografie**


- Alderson (2007). “The CEFR and the need for more research”. *The Modern Language Journal*, 91(4): 659–663.
- Alderson & Banerjee (2002). “Language testing and assessment (Part 2)”. *Language Teaching*. 35: 79-103.
- Alderson, Clapham & Wall (1996). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.
- Bachman (2000). “Modern language testing at the turn of the century assuring that what we count counts”. *Language Testing*. 17(1): 1-42.
- Bachman & Palmer (2010). *Language Assessment in Practice*, Oxford: Oxford University Press.
- Brindley (2001). “Language assessment and professional development”. In Elder, Brown, Hill, Iwashita, Lumley, McNamara & O’Loughlin (Eds.) (2001). *Experimenting with uncertainty: Essays in honour of Alan Davies*, Cambridge: Cambridge University Press: 126–136.
- Cassady & Johnson (2002). “Cognitive test anxiety and academic performance”. *Contemporary Educational Psychology*, 27 (2): 270-295.
- Chapelle, Jamieson & Hegelheimer (2003). “Validation of a web-based ESL test”. *Language Testing*. 20(4): 409-439.
- Fulcher (2012). “Assessment Literacy for the Language Classroom”. *Language Assessment Quarterly*, 9(2): 113-132.
- Inbar-Lourie (2008). “Constructing a language assessment knowledge base: A focus on language assessment courses”. *Language Testing*, 25(3): 385–402.
- Jones, N. (2011). “Defining an inclusive framework for languages”. Plenary talk at *ALTE 4th International Conference*, Kraków: Polen.
- Knoch (2009). “Diagnostic assessment of writing: A comparison of two rating scales”. *Language Testing*, 26(2): 275–304.
- Lantolf (2009). “Dynamic assessment: The dialectic integration of instruction and assessment”. *Language Teaching*, 42(3): 355–368.
- Lumley (2002). “Assessment criteria in a large- scale writing test: What do they really mean to the raters?”. *Language Testing* 19(3): 246–276.
- McNamara & Roever (2006). *Language testing: The social dimension*. London: Blackwell.
- McNamara (2008). *Language Testing*. Oxford: Oxford University press.
- McNamara (2006). “Validity in Language Testing: The Challenge of Sam Messick’s Legacy”. *Language Assessment Quarterly*. 3(1): 31-51.
- Shepard (2000). “The role of assessment in a learning culture”. *Educational Researcher*, 29(7): 4–14.
- Struyf (2004). *Evalueren: een leerkans voor leraren en leerlingen*. Leuven: Universitaire Pers Leuven.
- Weigle (1994). “Effects of training on raters of ESL compositions”. *Language Testing*, 11: 197-223.